AD_____

GRANT NUMBER DAMD17-98-1-8018

TITLE: False-Negative Interpretations in a CAD Environment

PRINCIPAL INVESTIGATOR: Bin Zheng, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: July 1999

TYPE OF REPORT: Annual

PREPARED FOR: Commanding General
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1999	Annual (1 Jul 98 - 30 Jun 99)

4. TITLE AND SUBTITLE
False-Negative Interpretations in a CAD Environment

5. FUNDING NUMBERS
DAMD17-98-1-8018

6. AUTHOR(S)
Bin Zheng, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this project is to examine the impact of CAD schemes on the diagnostic performance of radiologists, in particular, the change of false-negative interpretations under a CAD cueing environment. Based on the proposed schedule of this project, we have completed the selection of 120 subtle cases that are used in the observer performance study. All the images have been processed, and the sensitivity of the CAD schemes has been adjusted to generate different cueing levels in each image. We have designed and implemented an automatic image display system. This computer-controlled system has the capability of randomizing case selection for each reading session and to record the diagnostic results. After finalizing the study protocol and performing pre-study training for participating radiologists, the main reading experiment is now underway. The radiologists began reading cases in May. Once the readings are completed, data analyses will be performed using ROC-type methodology.

14. SUBJECT TERMS
Breast Cancer, computer-assisted diagnosis in mammography, cueing, false-negative interpretations, false-negative findings, observer performance studies, ROC analysis

15. NUMBER OF PAGES
31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

__√__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Bin Zhou 7/1/99

PI - Signature Date

TABLE OF CONTENTS

Front Cover	1
Standard Form 298	2
Foreword	3
Table of Contents	4
Introduction	5
Body (Statement of Work Tasks)	5
1. Case Selection	5
2. Case Preparation	6
a) CAD processing and cueing mode design	6
b) Implementation of a computer-controlled image display system	7
3. Finalizing study protocol and performing pre-study training	7
4. Performing the main reading experiment	8
Key Research Accomplishments	8
Reportable Outcomes	8
Conclusions	8
References	N/A
Appendices	N/A

INTRODUCTION:

During the last decade, interest in computer-assisted diagnosis (CAD) schemes for the early detection of breast cancer on mammograms has been rapidly increasing, and a large variety of schemes have been developed and tested. As a result, it is believed that eventually CAD schemes could provide radiologists with useful information to improve the efficiency and accuracy in the diagnosis of breast cancer. However, prompting potential areas of abnormalities can affect the mammographic interpretation process, and unfortunately, the effect may not be always beneficial. Therefore, to better understand the radiological interpretation process, we are conducting this experiment to examine how different CAD cueing environments affect the error rate (particularly for false-negative interpretations). For this purpose, an observer performance experiment is being conducted using an ROC-type methodology. From the relationship between the CAD cueing levels and average diagnostic performance (i.e., areas under the ROC curves), we hope not only to better understand the impact of CAD cueing on diagnostic performance, but also demonstrate an optimal approach to use CAD schemes in the clinical screening environment.

BODY (Statement of Work Tasks):

During the first year of this project, we have performed the following tasks:

1. Case selection.

To prepare for the observer performance study in this project, we selected 120 mammographic studies from a large clinical database available in our laboratory. These studies were acquired on 120 patients undergoing routine mammographic screening at three different medical centers. Of these 120 studies, 85 are abnormal (positive) and 35 are negative. The positive cases include a total of 38 verified microcalcification clusters (27 malignant and 11 benign) and 57 masses (39 malignant and 18 benign). Most of these studies include two images (the same view of the left and right breasts), but some (31) have only one image. Table 1 lists the number of studies in different categories. All positive cases were verified by pathological (biopsy) reports. All negative cases are determined based on follow-up mammographic examination results. All the studies were considered "subtle" ones, because the images involve either subtle abnormalities or complex, but normal anatomy. All the original film mammograms from these cases have been digitized in our laboratory using a high quality film digitizer with 12 bit gray-level resolution and 100 μm ×100 μm pixel sizes.

Table 1: Number of image studies in different categories. (M = malignant, B = benign).

	Mass		Microcalcification Cluster		Mass and Cluster		Negative	Total Cases
	M	B	M	B	M	B		
Single image studies	10	1	11	3	1	1	4	**31**
Two image studies	20	16	7	7	8	0	31	**89**
Total studies	30	17	18	10	9	1	35	**120**

2. Case Preparation

a) CAD processing and cueing mode design.

To find suspicious regions (for both masses and microcalcification clusters), every image in the database was first processed by our CAD scheme. This CAD scheme utilizes a rule-based classifier to detect microcalcification clusters and an ANN (artificial neural network) to identify mass regions. 38 true-positive clusters and 28 false-positive clusters (or 0.14 false-positive detections per image) were detected by the scheme. In order to include more false-positive clusters in the experiment, we opened (loosened) the rules except in the final re-clustering stage of the scheme. Then the false-positive clusters identified by the modified scheme were increased to 95 (or 0.46 per image). In mass detection, after image segmentation and multi-layer topographic region growth, a total of 57 true- and 774 false-positive regions (or 3.7 per image) were identified in these 209 images. The ANN was then used to classify these regions. The ANN assigned a score (from 0 to 1), which correlates with the likelihood of the region representing a true mass. All of the identified masses and microcalcification clusters described above were used as candidates for cueing during the observer performance study. In this experiment, five reading modes were designed as shown in Table 2. Thus, each observer will read each study five times under five different reading conditions.

Table 2: Five reading modes in the observer performance study.

Reading mode	ROI Cued	Cued sensitivity	Cued FP / image
1	No	0	0
2	Yes	0.9	0.5
3	Yes	0.9	2
4	Yes	0.5	0.5
5	Yes	0.5	2

In each mode, both masses and microcalcification clusters have the same cueing sensitivity. This way, we can not only evaluate the overall detection accuracy of the readers, but also examine whether the readers have different responses to the detection of either masses or microcalcification clusters under the same cueing sensitivity. There are a total of 95 verified abnormalities (57 masses and 38 microcalcification clusters) in the database. The selection of regions for cueing mass or clusters was performed independently. Each region is assigned a number. The cued regions were randomly selected. In modes 2 and 3 (see Table 2), the sensitivity level is 0.9. 51 true-positive mass regions and 34 true-positive clusters were selected and cued. In modes 4 and 5, the sensitivity is 0.5. 29 masses and 19 clusters are cued. Modes 2 and 4 average 0.5 false-positive identifications per image, while modes 3 and 5 average 2 false-positive identifications per image. To select the false-positive regions, we first included all 28 initially detected false-positive microcalcification clusters in modes 2 and 4. In modes 3 and 5, all 95 microcalcification clusters were selected and cued. Then, two threshold values were used to select false-positive mass regions. Using these thresholds, 79 mass regions were included in level one. Together with the 28 initial false-positive detections, modes 2 and 4 included a total of 107 false identifications (or 0.51 per image). 334 false-positive mass regions were selected for the second level. Together with the initial 95 false-positive clusters, reading modes 3 and 5 included a total of 429 false-positive regions (or 2.06 per image).

b) Implementation of a computer-controlled image display system.

The readings in this study are performed on soft display. The radiologists will read these 120 studies in a random order five times (five reading modes). We designed, tested, and implemented an automatic image display and control system. Each session includes a fixed number of 30 studies. A computer program randomly selects display order. Based on the reading mode in each session, the cueing areas will be appropriately marked. To reduce the bias due to remembering specific cases read before, the computer program excludes cases that had been read within a specified period into the current reading session. The radiologist can view two images side by side displayed on the monitor at a reduced resolution, or the radiologist can examine full resolution images, one at a time using scrollbars in both vertical and horizontal directions (zoom and scroll). The program is designed to accept radiologist's inputs. If a suspicious region is identified, the radiologist points the arrow to the center of the area and clicks the mouse. A message window then appears, followed by a confidence slider window for scoring purposes. A computer mouse is the only tool needed for the radiologists to input their diagnostic decisions.

3. Finalizing study protocol and performing pre-study training.

We have finalized the study protocol and selected seven radiologists to participate in this observer performance study. All selected radiologists are Board certified with a minimum of three years' experience in the interpretation of mammograms. We have written and tested a comprehensive "Instructions for Readers" document that is provided to each participating reader. A set of sample cases have also been selected and incorporated into our display system. These samples are used to train readers and familiarize them with the image display and diagnostic scoring system.

4. Performing the main reading experiment.

The main reading experiment is now under way in our laboratory. We expect that the reading experiment will be completed by April 2000. Once all the reading sessions are completed, we will analyze the data using ROC-type methodology.

KEY RESEARCH ACCOMPLISHMENTS:

- Selected 120 mammographic cases
- Designed prompting cues for observers participating in the study
- Incorporated the prompting cues in the CAD system
- Algorithm development to perform study on soft display
- Finalized study preparations
- Initiated radiologist training on workstation and prompting system
- Initiated the main reading experiment

REPORTABLE OUTCOMES:

1. Zheng B, Chang YH, Wang XW, Good WF, Gur D, Application of a Bayesian belief network in a computer-assisted diagnosis scheme for mass detection, *Proc SPIE on Medical Imaging* **1999;** 3661-167.

2. Zheng B, Wang XH, Chang YH, Good WF, Automatic detection of nipple and chest wall in digitized mammograms, *Proc Computer Assisted Radiology and Surgery, 13th International Symposium and Exhibition,* Paris, France, June 23-26, **1999.**

3. Zheng B, Good WF, Wang XH, Chang YH, Comparison of artificial neural network and Bayesian belief network in a computer-assisted diagnosis scheme for mammography, *Proc International Joint Conference on Neural Network,* Washington, DC, USA, July 10-16, **1999**

4. Zheng B, Good WF, Chang YH, Wang XH, Applying a genetic algorithm for the improvement of decision making in medical image diagnosis, *Proc IASTED International Conference on Artificial Intelligence and Soft Computing,* Honolulu, USA, August 9-12, **1999.**

CONCLUSIONS:

There are five research tasks listed in the Statement of Work of this project. In the first year, we have completed the first three tasks. Task four is now under-way. Thus, the study is moving forward according to the proposed plan. Due to the blind nature of this project, the hypothesis cannot be tested before readers complete all the reading sessions. Although at this current stage we have not published any data or statistical analysis of results that are directly related to this observer performance study, we presented several CAD related papers in four different international conferences, which acknowledge the support of this research grant [see Reportable Outcomes, 1-4].

Proceedings of SPIE's International Symposium
On Medical Imaging 1999
February 20-26, 1999 in San Diego, CA, USA

Application of a Bayesian Belief Network in a Computer-Assisted Diagnosis Scheme for Mass Detection

Bin Zheng, Yuan-Hsiang Chang, Xiao-Hui Wang, Walter F. Good, and David Gur

Radiological Imaging Division,
Department of Radiology,
University of Pittsburgh,
Pittsburgh, PA 15261, USA

ABSTRACT

The purpose of this study is to investigate the use of a Bayesian belief network (BBN) in a computer-assisted diagnosis (CAD) scheme for mass detection in digitized mammograms. In this study, two independent image sets were used. The training image set included 306 verified positive mass regions in 545 images (or 217 mass cases). The testing image set included 349 mass regions in 433 images (or 189 cases). All 978 images were first processed by our rule-based CAD scheme. After image segmentation and adaptive topographic region growth, 288 regions depicting verified masses and 2,204 suspicious but actually negative regions were identified in the training image set. In the testing set, 304 positive mass regions and 1,586 negative regions were identified. Fifty features were computed for each region. Then, BBN was constructed in order to classify these regions as positive or negative for mass. To optimize the number of active nodes in the BBN, a genetic algorithm (GA) was used to search for an optimal subset of features. Twelve GA selected local features and four additional global image-based features were then used to construct the BBN. The conditional probabilities across the BBN were computed using the regions identified from the training image set. The performance of the BBN was evaluated using an ROC methodology. To demonstrate the potential utility of the BBN, we compared the results using the BBN with that of an artificial neural network (ANN) with the same set of input features. The BBN achieved an area under the ROC curve (A_z) of 0.873 $\pm$ 0.009 in classifying the 304 positive and 1,586 negative regions in the testing set. The highest A_z value achieved by the ANN was 0.858 $\pm$ 0.012. After incorporating the BBN into our CAD scheme as the last classification stage, we detected 80% of 189 positive mass cases (in 433 testing images) with an average detection rate of 0.76 false-positive regions per image. Therefore, this study demonstrated that a BBN approach could yield a comparable performance to that using other classifiers. Using a probabilistic learning concept and interpretable topology, the BBN provides a flexible approach to improving CAD schemes.

Key Words: Computer-assisted diagnosis, Bayesian belief network, Artificial neural network, Digital Mammography, Cancer, Breast cancer diagnosis.

1. INTRODUCTION

After more than a decade of intensive research in the computer-assisted diagnosis (CAD) of mammography by a large number of groups, many schemes for the detection of masses and microcalcification clusters have been developed [1-10]. Although significant progress has been made in these new schemes, the result of a prospective clinical study demonstrated to date a significant reduction in performance as compared with that achieved using the databases for the optimization of the schemes [11]. Such results raise questions concerning feature domain coverage

when limited size databases are used for development of CAD schemes. The robustness of scheme performance depends on many factors, including but not limited to case difficulty [12], size of training database [13], and validation methods [14]. Data over-fitting during development is also a concern when CAD performance is evaluated using independent databases [15].

Artificial neural networks (ANN) [16-18] and decision trees (DT) [2,5,18] have been widely used in current CAD schemes to classify positive and negative regions. However, for ANN and DT based schemes data over-fitting is a primary issue [19]. Using a "hill-climbing" method to search for an optimal separation model (boundary) in a sparsely sampled multi-dimensional feature space, the classifier can easily be over-fitted. As a result, with increase in the number of features or training iterations, CAD schemes that utilize limited size databases are likely to perform poorly during independent testing. In contrast, the Bayesian Belief Network (BBN) uses a probabilistic approach to determine an optimal segmentation given a specific database [19]. Because of this approach, the BBN method has attracted wide research interests in several machine learning areas [19,20]. It has also been tested in a limited manner for the computer-assisted diagnosis of breast cancer, using the information from radiologist's reports in conjunction with data from physical examination and patient clinical history [21]. However, BBN based schemes have not been incorporated into CAD to identify masses in digitized mammograms.

The motivation for this preliminary study was to investigate the potential utility of a BBN in CAD schemes for ass detection. A description of the approach, as well as the preliminary experimental results, is provided here.

2. MATERIALS AND METHODS

2.1. Clinical database

Two independently acquired image sets were used in this study. One was used for training and the other for testing of the scheme. The training set included 545 images acquired on women undergoing mammographic procedures at the University of Pittsburgh Medical Center and its affiliate hospitals and clinics in Pittsburgh, PA. The testing set included 433 images provided to us by a research group from Washington University Medical School in St. Louis, MO. All 978 images were digitized in our laboratory using a laser film digitizer (Lumisys 150) with a pixel size of 100 μm × 100 μm and 12-bit gray-level resolution. The digitized images were then sub-sampled by a factor of four in both directions to generate new images of approximately 600 × 450 pixels. 306 and 349 visible mass regions are included in the training and testing databases, respectively. The locations of all these mass regions were marked by expert radiologists on the original mammograms. Because in some cases masses are only visible on one view (either mediolateral oblique (MLO) or cranio-caudal (CC)) images and in other cases only one view image was acquired, the number of actual mass cases are 217 and 189 depicted in the training and testing databases, respectively.

The 978 images were individually processed by our multi-layer topographic based CAD scheme which has been described elsewhere [4]. In brief, this scheme has three distinct stages to detect mass regions in a digitized mammogram. The first stage of image segmentation (including dual kernel filtering, subtraction, thresholding, and labeling) is used initially to search for all suspicious regions (approximately 20 regions per image in these two databases). Based on local contrast measurement, the second stage uses an adaptive region growth algorithm to define three topographic layers for each suspicious region. In each growth layer, a set of simple intra-layer boundary conditions on growth ratio and shape factor of the region is applied to eliminate a large number of initial suspicious regions. After the second stage, the number of suspicious regions (including actually positive and negative regions) decreases to 4,382 (or approximately 4.5 per image) when applied to these 978 images. For each of remaining regions, a set of features is automatically computed. In the third stage of the scheme, a nonlinear multi-layer multi-feature analysis is applied to classify positive and negative regions. The classification tools that have been previously tested for this purpose in our studies include a rule-based classifier [4], an ANN [13], and set enumeration trees [18]. In this study, a BBN was incorporated as a classification tool. All 4,382 suspicious regions identified by the second stage of the scheme were included in the study to develop and test the BBN classifier.

As a result of this selection process, 288 actually positive regions and 2,204 suspicious but actually negative regions were identified in the training database after the second stage of our CAD scheme. In the testing

database, 304 positive regions and 1,586 negative regions were identified. The diagnostic "difficulty" of the 592 positive and 3,790 negative regions, as represented by conspicuity or "lesion contrast" divided by "surrounding complexity" [22], have been reported elsewhere [23].

2.2. Topology of a Bayesian belief network

A BBN is a graphical data structure that compactly represents the joint probability distribution of a problem domain by exploiting conditional dependencies, and it captures knowledge of a given problem domain in a natural and efficient way [24]. A BBN builds an "*acyclic*" graph in which nodes represent feature variables, and connections between nodes represent direct probabilistic influences between the variables. Due to the properties of "*acyclic*" connection and *d*-separation defined in the BBN [25], there is no feedback loop between any nodes and the lack of connection (or path) between two nodes indicates the probabilistic independence of two variables. Each node in a BBN represents one feature variable. Each variable must have two or more discrete states. For a discrete variable, its digital or symbolic values can be used as the states of the node. For a continuous variable, the values must be segmented into discrete states. Each state is then associated with a probability value; for each node, the summation of probability values for all states equals to one. The conditional probabilities between connected nodes can be assigned by established statistic data [21] or computed from a set of measured training data [26]. In general, when the network structure is given in advance and the variables are fully observable in the training examples, learning the *prior* and *conditional* probabilities is a straightforward procedure [25].

The nodes in the BBN were represented using the features computed by our CAD scheme. The feature set for each suspicious region contained both local and global features. The local features were the features computed from the interior of three topographic growth layers and surrounding background of each region [13,23]. The global features were extracted from the whole image [27]. In this experiment, 50 local and four global features were initially computed. The four global features were (1) the image view, (2) the region location, (3) ratio between peak value in the histogram and the size of breast area, and (4) the average local fluctuation of the breast tissue. Figure 1 demonstrates a topology of the BBN designed for this experiment. Although the four global features are largely independent to the presence of the suspicious mass regions, they may have impact on the detection difficulty of the regions. Thus, these features (x_{1i}, $i = 1,2,3,4$) were placed in the top of detection node, (Mass or Y in figure). Meanwhile, since all the local features were computed from a suspicious region, they were placed as child nodes (x_{3j}, $j = 1,2,\ldots,N$) of the detection node in the BBN.

Each node in the BBN must have at least two discrete states. The detection node has two states, *Positive* or *Negative*. The global features were discrete features. Feature 1 has two states (*MLO* or *CC*), feature 2 has four states (*upper outer, upper inner, lower outer,* and *lower inner*), features 3 and 4 have three states (*low, middle,* and *high*). The local features are continuous numbers, which must be segmented into discrete states. Based on our experimental result, which has been reported elsewhere [28], all local features were segmented into five discrete states. Using the range of values for each, the segmentation boundaries were determined with the criterion that all states contained approximately the same number of regions.

Next, using our training database, all the required conditional probabilities used in the BBN were computed. Based on the number of possible permutation and combination, the following 72 conditional probabilities were required for a complete probability table between the global feature nodes and the detection node:

$$P_1(Y = Positive \mid x_{11} = 1, x_{12} = 1, x_{13} = 1, x_{14} = 1);$$
$$P_2(Y = Negative \mid x_{11} = 1, x_{12} = 1, x_{13} = 1, x_{14} = 1);$$
$$P_3(Y = Positive \mid x_{11} = 1, x_{12} = 1, x_{13} = 1, x_{14} = 2);$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots;$$
$$P_{72}(Y = Negative \mid x_{11} = 2, x_{12} = 4, x_{13} = 3, x_{14} = 3).$$

Because $P_2 = 1 - P_1$, ..., $P_{72} = 1 - P_{71}$, only 36 values are independent. The detection node also has up to 50 child nodes. Since each local feature has five states and detection node has two states, 10 conditional

probabilities for each node, such as $P(x_{31} = 1 | Y = Positive)$ and $P(x_{31} = 1 | Y = Negative)$, were computed. 8 of 10 conditional probabilities are independent in each node. Thus, ups to 400 independent conditional probabilities were determined between the detection node and local feature nodes. All of these conditional probabilities form a complete set of connection weights applied in the BBN.

2.3. Optimal feature selection using a genetic algorithm

Although 50 local features were initially computed for each suspicious region, many may be highly correlated or have little contribution to the separation of actually positive and negative mass regions. To reduce the redundancy, we used a genetic algorithm (GA) [29] to select an optimal feature set.

The GA software we used, *GENESIS*, was downloaded from Prime Time Freeware for AI [30]. The software provides a basic program structure for the optimization, such as initialization, evaluation, recombination, crossover, and mutation of chromosomes. To run the program, the user needs to provide both a fitness function and an evaluation criterion. The user also needs to run first its setup routine to set up the encoding format of the chromosomes as well as the other required parameters. In this experiment, each chromosome contained 50 genes, which represented 50 local features. The binary coding was used to create the chromosome, with *1* indicating the presence of a gene (the feature was used in the BBN) and *0* indicating absence of the gene (the feature was not used in the BBN). The initial population size of the chromosomes was selected as 50. In order to incorporate our previous experiences in the feature selection and also to achieve a diverse initial population, about one third of initial chromosomes (16 of them) were specifically selected with small number of bits of *1* ($\leq$ 10), while the rest of initial population was randomly assigned by software. The crossover rate, the mutation rate, and the generation gap were set up at 0.6, 0.001, and 1.0, respectively. These three values are default levels suggested by the *GENESIS*.

The evaluation subroutine in *GENESIS* was connected to the BBN testing. All conditional probabilities required in the BBN were computed for the 288 positive and 2,204 negative regions in the training database. The independent testing database that included 304 positive mass regions and 1,568 suspicious but negative regions was used in the GA optimization experiment. The chromosomes generated by the GA determined which features were selected for the BBN. The testing result from the BBN was analyzed by ROC methodology using the program ROCFIT [31]. The chromosomes that produced higher areas under the ROC curves (A_z values) survived and used to create new chromosomes in next generation. The GA was terminated when better chromosome could not be found in the new generation. Then, an optimal local feature set that produced the best A_z value was actually used in the BBN.

2.4. Comparison between the ANN and BBN the same CAD scheme

Since the performance of a CAD scheme depends heavily on the case difficulty and there is no commonly accepted method to measure this parameter [12], without an objective comparison, reporting the performance of the BBN in absolute terms may be meaningless. Therefore, we compared the performance of the BBN with an ANN. ANN performance in this case has been demonstrated in a large number of independent studies, including ours [13,18,23]. The ANN was trained and tested using the same features and the same databases as that used by the BBN. The detailed structure and training method of the specific ANN has been reported elsewhere [13]. To test the robustness of the ANN, five different iterations (500, 1.000, 2,000, 3,000, and 5,000) were used for training. After each training, the ANN was tested with the same independent testing database using ROC analysis.

2.5. Evaluation of a BBN-based CAD scheme for mass detection

Finally, we incorporated the BBN into our CAD scheme as its classification tool in the third stage. Areas under the ROC curves (A_z) and false-positive detection rates at 80% detection sensitivity were used as summary indices of performance.

3. RESULTS

After evolution of 50 generations, the GA selected 12 local features from the original 50 features. The 12 features were listed in table 1 and their methods of computation have been reported elsewhere [13,23,27]. Using these 12 features the area under ROC curve (A_z) for the testing database was 0.873 ± 0.009.

Table 1: Local feature set selected by the GA from the 50 original features.

Feature Number	Description of the Feature
1	Region size in the third growth layer.
2	Region contrast in the third layer.
3	Skewness of pixel values inside the third growth layer.
4	Circularity in the third layer.
5	Size growth ratio between the second and the third layers.
6	Ratio of local minimum pixels inside the growth region of the third layer.
7	The central position shift between the second and the third layer.
8	Ratio of pixels whose values are smaller than growth threshold in the third layer in surrounding background.
9	Region conspicuity.
10	Standard deviation of pixel values in the second layer.
11	Circularity of growth region in the first layer.
12	Ratio of largest and shortest radial lengths in the first layer.

With the same 16 features (12 local ones selected by the GA and 4 global ones), an ANN, which involved 16 input neurons and 8 hidden neurons, was trained and tested using the same databases. Figure 2 demonstrates the over-fitting pattern of the ANN. The optimal testing performance of the ANN is 0.858 ± 0.012 using 1,000 training iterations. As a result, the BBN outperformed the ANN in testing by 1.5%. Figure 3 shows the two ROC curves that representing the highest performance achieved using the BBN and the ANN in this experiment.

After incorporating the BBN into our CAD scheme, the performance on the complete 433 images in the testing database is demonstrated in figure 4. There are total 189 mass cases (or 349 visible mass regions) in this database. Before using the BBN, the CAD scheme (the first two stages) has detected 180 mass cases (95.2%) or 304 mass regions (87.1%) with average 3.6 false-positive regions per image. Figure 4 shows two curves. One curve represents the result of case-based detection, where a mass is considered as detected by the scheme if it is detected in either one view (CC or MLO) or both views. The second curve (dash curve) shows the result for a region-based detection, where one mass depicted in two view images is considered as two independent detection targets. For example, by setting appropriate thresholds in these two curves, the CAD scheme can detect 80% of 189 positive cases with an average of 0.76 false positive regions per image, or 80% of 349 positive regions with an average of 1.45 false positive identifications per image. At a case-based detection sensitivity level of 80%, 57.5% (or 249) images did not have any false-positive regions identified.

4. DISCUSSIONS

To develop a successful CAD scheme, it is important to have both high performance and robustness in independent testing with "images never seen" to the scheme. Although it has been a popular classification tool for the CAD in mammography, ANN has several disadvantages in classifying complex and diverse data, in particular,

when the number of training samples is limited. First, in a complex and noisy multi-dimensional feature space, the ANN may reach a locally optimal solution based on the randomized selection of initial values of the weights. In such a case, the ANN may become unstable. Second, an ANN uses a "black-box" learning approach that makes ANN's knowledge inaccessible to simple human understanding. There is no way to meaningfully explain the reasoning of an ANN and the weights associated with it [32]. Therefore, it is difficult to select a best topology of the ANN for a large number of specific applications. Third, over-fitting is also a significant difficulty in ANN training. Due to these factors, it is sometime difficult to maintain a robust performance of a CAD scheme using this methodology when the ANN is trained by a small number of samples. Robustness issues are not unique to ANN methodology, it also exists in other optimization algorithms (i.e., rule-based discriminant functions and decision-trees). As a result, performance of CAD schemes may be significantly variable not only when testing a new set of images but also when testing the same set of images acquired under different condition (e.g., digitized at different times [33] or when using different digitizers [34]).

The BBN uses a different learning approach as compared with the ANN. Unlike "a black box" approach, BBN can more flexibly represent incomplete knowledge or uncertainty. It allows users to specify dependence and independence of features in a more natural way through the network topology. The learning process in a BBN to determine the weights (the conditional probabilities) between connected nodes can be directly computed from actual measurements. The meaning of the weights is also more understandable [35]. By eliminating "hill-climbing" learning type process, the danger of over-fitting can be reduced in this approach. The main factor that affects the robustness of the network originates from the bias of the learning samples, which affects the accuracy of the conditional probabilities.

In summary, ANN and BBN are two common machine learning algorithms for pattern recognition [19]. Unlike ANN, BBN has not been applied to the mass detection in mammograms to date. The results of our preliminary experiment are encouraging. BBN may provide a flexible, stable, and understandable approach to improving the performance and robustness of the CAD schemes.

5. ACKNOWLEDGMENTS

The authors thank William Reinus, MD, and the research group at Washington University, St. Louis, MO, for providing us with some of images used in this study. This work was supported in part by National Cancer Institute under grants CA77850, CA82912, and US Army under grant DAMD17-98-1-8018.

6. REFERENCES

1. C.J. Vyborny, M.L. Giger, "Computer vision and artificial intelligence in mammography," *Am J. Roentgen*, 162: 699-708, 1994.
2. W.P. Kegelmeyer, J.M. Pruneda, P.D. Bourland, A. Hills, M.W. Riggs, M.L. Nipper, "Computer-aided mammographic screening for speculated lesions," *Radiology*, 191: 331-337, 1994.
3. B. Zheng, Y.H. Chang, M. Staiger, W. F. Good, D. Gur, "Computer-aided detection of clustered microcalcifications in digitized mammograms," *Acad Radiol*, 2: 655-662, 1995.
4. B. Zheng, Y.H. Chang, D. Gur, "Computerized detection of masses from digitized mammograms using a single image segmentation and a multi-layer topographic feature analysis," *Acad Radiol*, 2: 959-966, 1995.
5. H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, R.A. Clark, "Markov random field for tumor detection in digital mammography," *IEEE Trans Med Imaging*, 14: 565-576, 1995.
6. F. Lefebvre, H. Benali, "A fractal approach to segmentation of microcalcifications in digital mammograms," *Med Phys*, 22: 381-390, 1995.
7. Y.H. Chang, B. Zheng, D. Gur, "Computerized identification of suspicious regions for masses in digitized mammograms," *Invest Radiol*, 31: 146-153, 1996.

8. D. Wei, H.P. Chan, N. Petrick, M.A. Helvie, D.D. Adler, M.M. Goodsitt, "False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis," *Med Phys*, 24: 903-914, 1997.
9. W. Zhang, H. Yoshida, R.M. Nishikawa, K. Doi, "Optimally weighted wavelet transform based on supervised training for the detection of microcalcifications in digital mammograms," *Med Phys*, 25: 949-956, 1998.
10. A.J. Mendez, P.G. Tahocas, M.J. Loda, "Computer-aided diagnosis: automatic detection of malignant masses in digital mammograms," *Med Phys*, 25: 957-964, 1998.
11. R.M. Nishikawa, M.L. Giger, R.A. Schmidt, D.E. Wolverton, B.S. Collins, K. Doi, "Computer-aided diagnosis in screening mammography: detection of missed cancers," (abstr), *Radiology*, 209(P): 256, 1998.
12. R.M. Nishikawa, M.L. Giger, K. Doi, "Effect of case selection on the performance of computer-aided detection schemes," *Med Phys*, 21: 265-269, 1994.
13. B. Zheng, Y.H. Chang, W.F. Good, D. Gur, "Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme," *Acad Radiol*, 4: 497-502, 1997.
14. G.D. Tourassi, C.E. Floyed, "The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis," *Med Decis Making*, 17: 186-192, 1997.
15. H.P Chan, B. Sahiner, R.F. Wagner, N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," *Proc SPIE*, 3338: 845-858, 1998.
16. H.P. Chan, B. Lo, B. Sahiner, K.L. Lam, M.A. Helvie, "Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network," *Med Phys*, 22: 1555-11567, 1995.
17. W. Zhang, Doi K, Giger ML, RM Nishikawa, R.A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," *Med Phys*, 23: 595-601, 1996.
18. R. Rymon, B. Zheng B, Y.H. Chang, D. Gur, "Incorporation of a set enumeration tree-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection," *Acad Radiol*, 5: 181-187, 1998.
19. T.M. Mitchell, *Machine Learning*, WCB McGraw-Hill, Boston, MA, 1997.
20. D. Michie, D.J. Spiegelhalter, C.C. Taylor, *Machine learning, neural and statistical classification*, (edited collection), Ellis Horwood, New York, NY, 1994.
21. C.E. Kahn, L.M. Roberts, K.A. Shaffer, P. Haddawy, "Construction of a Bayesian network for mammographic diagnosis of breast cancer," *Comput. Biol. Med*, 27: 19-29, 1997.
22. H.L. Kundel, G. Revese, "Lesion conspicuity, structure noise, and film reader error," *Am J. Roentgen*, 126: 1233-1238, 1977.
23. B. Zheng, Y.H. Chang, W.F. Good, D. Gur, "Assessment of mass detection using tissue background information as input to a computer-assisted diagnosis scheme," *Proc SPIE*, 3338: 1547-1555, 1998.
24. P.R. Harrison, J.G. Kovalchik, Expert system and uncertainty, in The handbook of applied expert systems, edited by Liebowitz J, chapter 8, CRC Press, Boca Raton, FL, 1997.
25. F.V. Jensen, *An introduction to Bayesian network*, Springer Verlag, New York, NY, 1996.
26. X.H. Wang, B. Zheng, W. F. Good, Y.H. Chang, "Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network," *International J. Medical Informatics*, in press, 1999.
27. B. Zheng, Y.H. Chang, D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms," *Acad Radiol*, 3: 806-814, 1996.
28. B. Zheng, Y.H. Chang, X.H. Wang, W.F. Good, D. Gur, "Feature selection for computerized mass detection in digitized mammograms using a genetic algorithm," *Acad Radiol*, in press, 1999.
29. G. Hluck, *Genetic algorithms*, in chapter 12 of *The handbook of applied expert systems*, edited by Liebowitz, J, CRC Press, Boca Raton, FL, 1997.
30. M. Kantrowitz, Prime time freeware for AI, Issue 1-1, Selected materials from the Carnegie Mellon University Artificial Intelligence Repository, 1994.
31. C.E. Metz, H.B. Kronman, P.L. Wang, J.H. Shen, "ROCFIT: A modified maximum likelihood algorithm for estimating a binormal ROC curve from confidence-rating data," Chicago: University of Chicago, 1985.
32. J. Diederich, "Explanation and artificial neural networks," *Int. J. Man-Machine Stud*, 37: 335-341, 1992.
33. Y.H. Chang, B. Zheng, D. Gur, "Robustness of computerized identification of masses in digitized mammograms: a preliminary assessment," *Invest Radiol*, 31: 563-568, 1996.
34. R.P. Velthuizen, L.P. Clarke, "Digitized mammogram standardization for display and CAD," *Proc SPIE*, 3335-20, 1998.
35. P. Haddawy, J. Jacobson, C.E. Kahn, "Generating explanations and tutorial problems from Bayesian networks," *Proc Annu Symp Comput Appl Med Care*, 770-774, 1994.

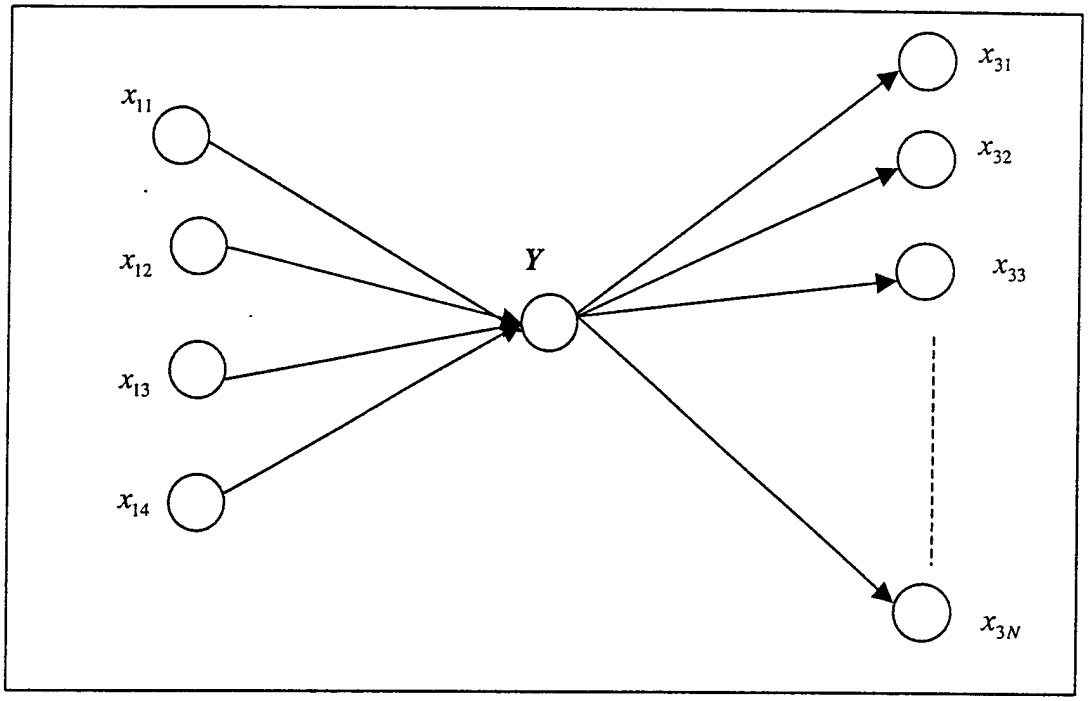

Figure 1: Topology of a Bayesian belief network (BBN) for mass detection used in this experiment.

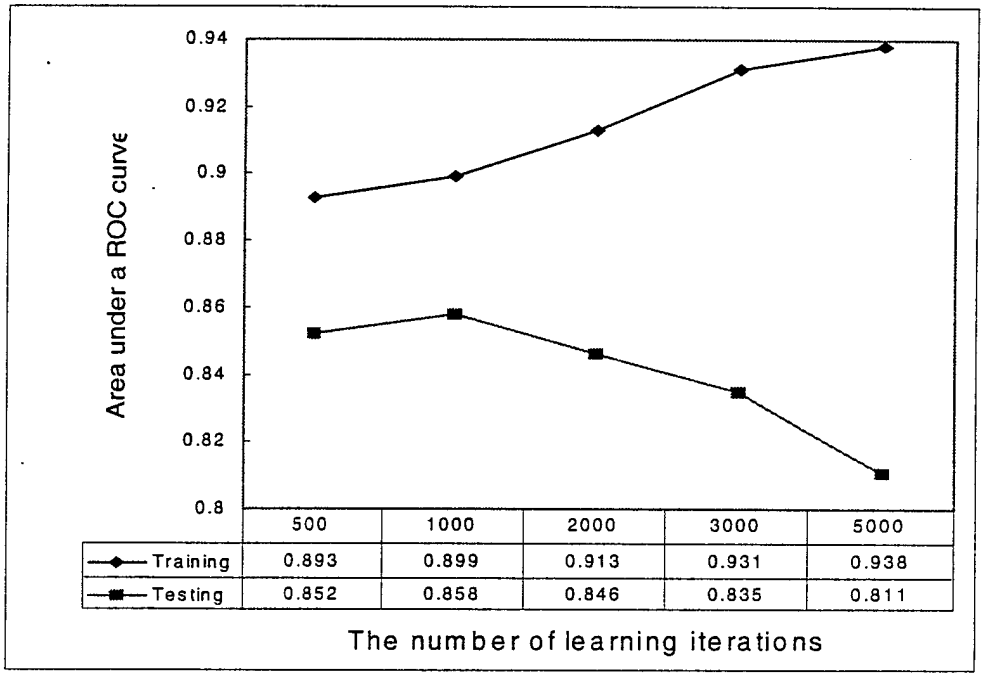

Figure 2: The distribution of A_z value using the ANN after different numbers of learning iterations. Standard deviations of the A_z values ranged from 0.007 to 0.013.

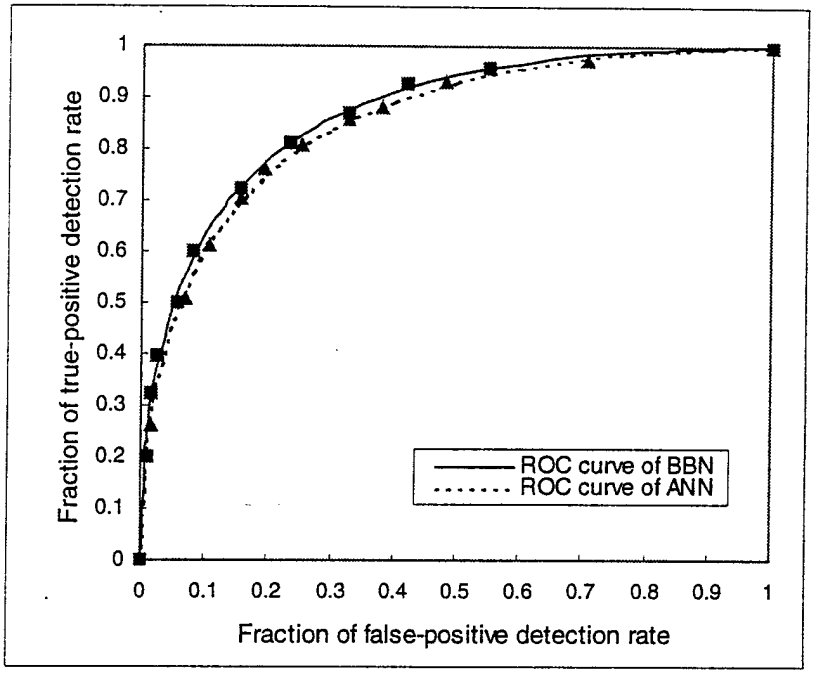

Figure 3: Comparison of two highest performing ROC curves using the BBN and the ANN. The A_z value for the BBN is 0.873 ± 0.009, and the A_z value for the ANN is 0.858 ± 0.012.

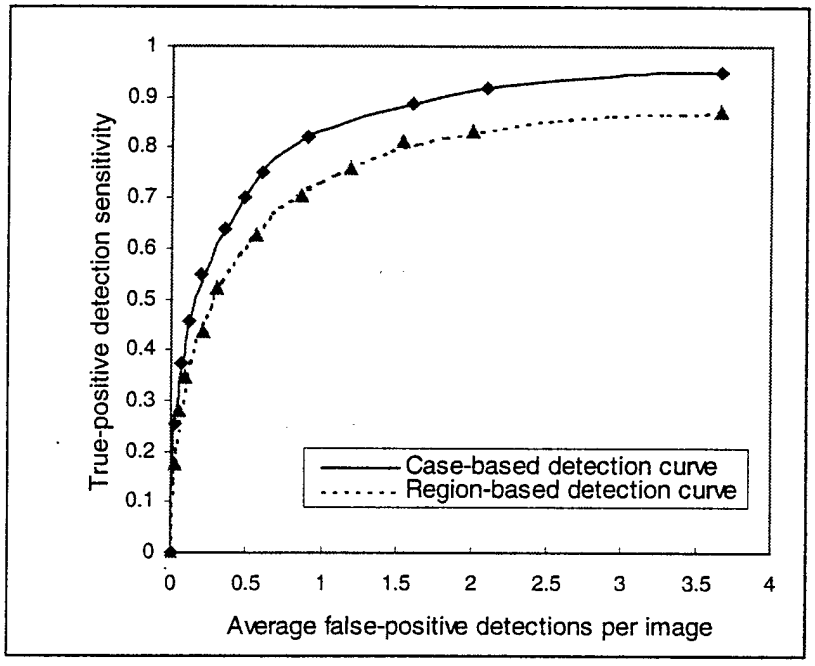

Figure 4: Case and Region-based performance of the complete CAD scheme on a testing with 433 images after incorporation of a BBN classifier.

Proceedings of 13th International Congress and Exhibition
On Computer Assisted Radiology and Surgery
June 23-26, 1999 in Paris, France

Automatic detection of the nipple and chest wall in digitized mammograms

Bin Zheng, Xiao-Hui Wang, Yuan-Hsian Chang, and Walter F. Good
Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15261, USA

This study developed a fully automatic algorithm for the detection of nipple and chest wall in digitized mammograms. The algorithm involves four steps to sequentially search for skin-air interface, chest wall, and nipple location. 334 images were used in the experiment. These images were divided into three difficult groups. Using the maximum matching difference of 10 mm between visual and automatic detections as a criterion, 99%, 82%, and 62% of nipple detections were matched in the easy, moderately difficult, and difficult image groups, respectively. For chest wall detection, 140 cases in a total 156 MLO images (90%) were matched. This study demonstrates a simple and fully automatic algorithm that has the potential to be applied in computer-assisted diagnosis (CAD) schemes of mammography.

1. INTRODUCTION

For the last decade, large number of CAD schemes in mammography has been reported, but most of them have not been successful in the clinical environment. One reason may be the different approaches used between radiologists and CAD schemes. The radiologists routinely compare corresponding regions of two images, either images of right and left breasts or images from two views of the same breast, to detect abnormalities. For a difficult case, radiologists may call images from previous examination for comparison. Thus, extraction and comparison of features from two related images might also be an important step to significantly improve CAD performance. The success of this approach relies on the correct registration of two images. Due to dominated soft tissue in the breast, the nipple is the only landmark. Developing an algorithm for automatic detection of nipple location is the first step in computerized image registration and feature comparison. Therefore, in this study, we developed a fully automatic and multi-stage algorithm to detect the nipple and chest wall (for MLO images). The accuracy of the algorithm was then evaluated by a large database. The detailed description of our algorithm, the database, and experimental results is reported here.

2. MATERIALS AND METHODS

The image database involved a total of 334 digitized mammograms, which were randomly collected from 92 women undergoing breast examinations at University of Pittsburgh Medical Center. These images were digitized in our laboratory using a digitization protocol reported before [1]. Due to the variety of image quality, the nipple locations may not be always visible in images. After analyzing the database, we set up following hypotheses:

1. A small and obvious protruding area in a smooth skin boundary (interface between breast and background) can indicate the location of a not in profile nipple.
2. If a nipple is in profile, it is likely to be located in an area where the pixel values are not only relatively unchanged but also significantly smaller than that of any other tissue regions near the skin boundary.

3. For images with poor visual quality around skin boundary, a point in skin boundary that has the longest distance to the chest wall is assumed as the location of a nipple.

Based on these hypotheses, we developed a multi-stage automatic algorithm to detect the nipple and chest wall in the image. It includes four major steps.

The first step is to find the skin-air interface in an image. In this algorithm, we use an iterative thresholding method to search for the smoothest transition boundary between the skin-air interface in the image. Typically, there are two major hills in an image histogram. One widely spread hill represents the pixel value distribution of breast area, while the another narrow hill in the higher digital value region indicates the air background. The threshold value to segment the skin-air interface is located in the valley between two hills. In this algorithm, a set of threshold values is selected in the valley. In each threshold, the computer program defines a segmented image and tracks the skin-air interface boundary. Then, the smoothness of all tracked boundaries is compared. The computation is performed based on the standard deviation of the distance in two adjacent points in the tracked skin line. The smoothest curve in this set of iterative tracking curves is selected as the skin boundary.

The second step is to search for a small but obvious protruding area along the skin boundary. If the nipple is visible by raising window level and reducing window size in the image, there will be a small but obvious protruding area outside the skin boundary. To find such a small protruding area, for every tracked point from in skin boundary, a line is drawn (or calculated) to link two points that are 40 track points away. The area covered by the tracked skin boundary and the line is computed. The maximum area detected along the skin boundary is the first candidate for the nipple. If this area is larger than a pre-determined value, the most protruding point in this area is identified as the nipple location.

The third step is to search for a small area with a substantially low pixel value but relative uniform distribution, which is located adjacent to the skin boundary inside the breast area. If there is no obvious protruding area in skin boundary as described in the previous step, the computer program is going to search for the nipple in profile. A square window of 20×20 pixels is used to scan along the skin boundary. The computer program measures the medium pixel value and the standard deviation of pixel values inside the window. Then, an area with the smallest medium pixel value is considered as the second candidate for the nipple. If it can pass a simple rule-based criterion, the area is considered to represent a nipple in profile and the center of the area is defined as the nipple location.

The fourth step is to detect chest wall. If the nipple is invisible and there is no clear skin-air interface in the image due to a variety of clinical reasons, the nipple position can not be detected in steps 2 and 3. Thus, we use the concept of maximum height of the breast border [2] to estimate the nipple location. The maximum height of the breast border in our algorithm is defined as the maximum distance between the skin boundary and chest wall. For the CC or a few LO images, we assume that the chest wall is parallel to the edge of the film, because chest walls are not visible in these images. Hence, the maximum distance is the same as the maximum height of the breast border. For the MLO images, the algorithm detects chest wall based on the process of maximum gradient search along each horizontal line scanning and line fitting of the maximum gradient points using least square method. Every image in the database has been oriented so that the chest walls always locates in the left side

of the image. Then from the top of the image the computer program scans horizontally from the left edge of the image until the 10 pixels before reaching the skin boundary. Along the scanning line, the computer program calculated the gradient of pixel value change along the line. The point with the maximum gradient ($g(x_i)_{max} > 100$) is considered as a point located in the chest wall for this scanning line. If in a scanning line, $g(x_i)_{max} \leq 100$, this line is skipped because there is no clear point that indicates the location of chest wall in this line. The computer program will automatically stop scanning when $g(x_i)_{max} > 100$ and $x_i < 5$. Then, the least square method is used to fit all the points (x_i, y_i) recorded as the maximum gradient point. The fitted line is defined as the chest wall. Once the chest wall is defined, the distance between the chest wall and every point in the skin boundary is computed. Then, the point in the skin line with maximum distance to chest wall is estimated as the nipple location. Correctly detecting the chest wall in the MLO images is also very useful for the future image registration and comparison. Thus, in this study, we compute the chest wall for all MLO images and examine the detection accuracy of the algorithm.

Since the performance of CAD schemes depends on the difficulty of the database [3], we set up three criteria to divide 334 images into three groups, which are easy, moderately difficult, and difficult groups. The easy group contains images where the image has a relatively clear skin boundary and the nipple location is visible. The moderately difficult group involves the images where the skin boundaries are vague and the nipple locations are ambiguous. The difficult group includes images where both the skin-air interface is difficult to separate and the nipple is totally invisible. Based on these criteria, the easy, moderately difficult, and difficult groups contain 139, 67, and 128 images, respectively. All of the nipple and chest wall locations were first visually marked or estimated if the nipple is invisible in the image. The coordinates for all the nipples and chest walls were saved in a "truth" file. Before testing the performance of the algorithm, 30 images from the easy group were randomly selected as training (rule-setting) images. Based on the analysis of these images, two simple identification criteria were set up in the algorithm. They are the minimum size of the protruding area and the minimum pixel value difference between skin boundary and low uniform density area. The remaining 304 images were then used to test the accuracy of the algorithm by comparing matching difference between the visually and automatically located nipple and chest wall of the same image. In this experiment, if the distance between two matching targets was smaller than 10 mm, two targets were considered as matched.

3. RESULTS

For the nipple detection in 304 testing images, 242 images (80%) were matched. In the three difficult image groups, 108 images matched (99%) in the easy group; 55 images (82%) matched in the moderately difficult group; and 79 images (62%) matched in the difficult group. Figure 1 demonstrates the matching histogram in these three groups.

For chest wall detection, the matching difference is smaller than 5mm in 125 images. In another 15 images the difference is between 5 to 10mm. As a result, 90% of cases (in 156 MLO images) are considered matched using the threshold of maximum 10mm difference.

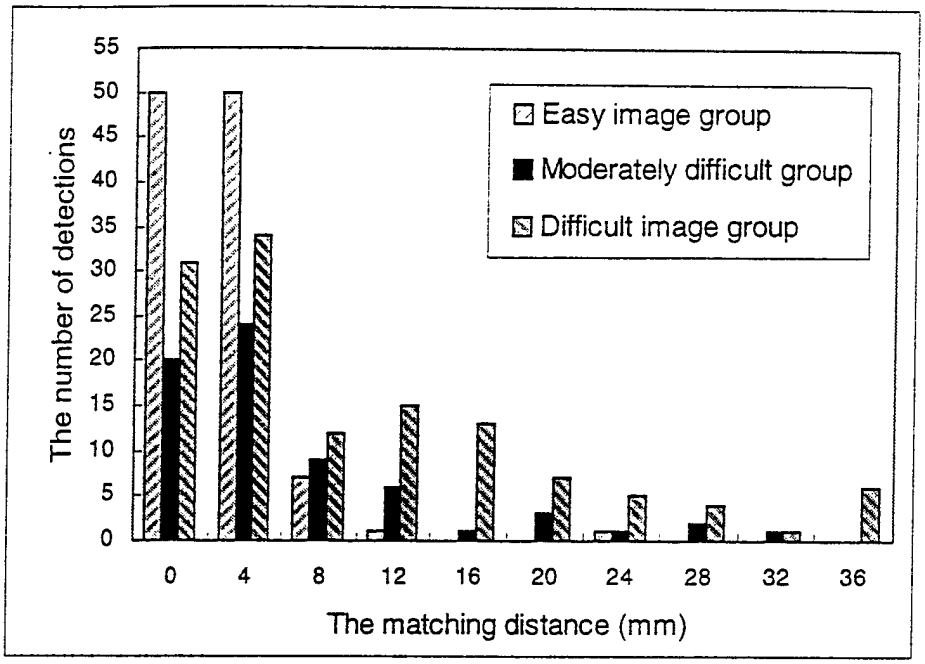

Figure 1: The histogram of matching distance in three image groups.

4. DISCUSSIONS

In this study, we developed a simple multi-stage algorithm to automatically detect the nipple and chest wall in digitized mammograms. The algorithm was tested by a relatively large and diverse database. The result is encouraging. For the cases where nipple positions can be visually located, the algorithm achieved very high detection accuracy. For other cases where nipple positions are invisible, using the maximum distance between skin line and chest wall as an estimation criterion also yielded a reasonable accuracy. It should be noted that for these invisible cases, visually located nipple positions might not be always accurate. Further improving our algorithm in the design and testing will be conducted in our future studies.

5. ACKNOWLEDGMENTS

This work was supported in part by National Cancer Institute (USA) under grants CA 77850, CA82912, and US Army under grant DAMD 17-98-1-8018.

6. REFERENCES

1. B. Zheng, YH Chang, and D. Gur, Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis, *Acad Radiol,* Vol. 2 (1995) 959-966
2. R. Chandrasekhar and Y. Attikiouzel, A simple method for automatically locating the nipple on mammograms, *IEEE Trans Med Imag*, Vol. 16 (1997) 483-494.
3. R.M. Nishikawa, M.L. Giger and K. Doi, Effect of case selection on the performance of computer-aided detection schemes, *Med Phys*, Vol. 21 (1994) 265-269.

Proceedings of the IJCNN International Conference
On Neural Networks
July 10-16, 1999 in Washington, DC, USA

Comparison of Artificial Neural Network and Bayesian Belief Network in a Computer-Assisted Diagnosis Scheme for Mammography

Bin Zheng, Yuan-Hsiang Chang, Xiao-Hui Wang, Walter F. Good
Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15261-0001
Bzheng@radserv.arad.upmc.edu

Abstract

Artificial neural networks (ANN) have been widely used in computer-assisted diagnosis (CAD) schemes as a classification tool to identify abnormalities in digitized mammograms. Because of certain limitations of ANNs, some investigators argue that Bayesian belief network (BBN) may exhibit higher performance. In this study we compared the performance of an ANN and a BBN used in the same CAD scheme. The common databases and the same genetic algorithm (GA) were used to optimize both networks. The experimental results demonstrated that using GA optimization, the performance of the two networks converged to the same level in detecting masses from digitized mammograms. Therefore, in this study we concluded that improving the performance of CAD schemes might be more dependent on optimization of feature selection and diversity of training database than on any particular machine classification paradigm.

Key Words: Artificial neural network, Bayesian belief network, Genetic algorithm, Computer-assisted diagnosis.

I. Introduction

After intensive investigation for more than a decade by many research groups, a large number of computer-assisted diagnosis (CAD) schemes in mammography have been developed [1-10]. For mass detection, the CAD schemes usually adopted three steps to identify positive and negative mass regions. In the first step, they use different image segmentation methods to filter out basic tissue structure and select initial suspicious regions. The second step is to compute or extract features from each suspicious region. Then, the third step uses a classifier to identify positive and negative regions based on the set of extracted features. Although each scheme was independently developed using databases of limited size, most current CAD schemes yield similar performance (i.e., 85% to 90% sensitivity with 1 to 2 false-positive regions per image in mass detection). Unlike many other pattern recognition problems where the feature domain is reasonably limited and well defined (e.g., optical character recognition), the feature space in CAD of mammography is very complex, due to wide diversity of normal tissue patterns and variety of abnormalities. To improve CAD performance, a large number of features are usually extracted. Most of these features are neither visible nor understandable by human observers. It is very difficult to find the correlation and effectiveness of these features in identifying masses in mammograms. Thus, many machine learning (classification) methods have been tested to identify the positive mass regions based on a set of computed features. Artificial neural networks (ANN) are the most popular paradigm used as a machine learning classifier in current CAD schemes.

The ANN uses "hill-climbing" approach to learn the correct response or output for each of the training samples. After training, the structure of the ANN has been self-organized to enable extrapolation when faced with new, yet similar, patterns, on the basis of "experience" with the training set. One attractive feature when using an ANN in a complex pattern recognition problem is that the required amount of *a priori* knowledge of the input features and internal system operation is minimal [11]. Although ANNs have the ability to learn complex patterns directly from observations, their reasoning process is inaccessible to human understanding and observers cannot be certain what the ANN has learned [12]. Because of this, it may be hard for physicians to accept and act on a computer system's advice without knowing the basis for the system's decision [13]. Furthermore, due to "hill-climbing" optimization process of the ANNs, the possible data over-fitting may

significantly deteriorate the robustness of ANN-based CAD schemes in real clinical environment [14].

Bayesian belief networks (BBN) use different training concept. A BBN is a causal probabilistic network that compactly represents the joint probability distribution of a problem domain by exploiting conditional dependencies. The BBN captures knowledge of a given problem domain in a natural and efficient way. A BBN can also explain its reasoning and can avoid the danger of data over-fitting [15]. Because of these unique characteristics, BBNs have been widely used in many machine learning applications [16]. In the area of computer-assisted diagnosis of breast cancer, some researchers have claimed that BBNs should perform much better and more reliably than ANNs [17]. Because in current CAD studies, different schemes have been trained and tested using different databases, performance of these schemes can not be compared [18].

In this study, we used the same database to train both an ANN and a BBN. After optimizing the topologies of these networks using a genetic algorithm (GA), an independent database was used to test the performance and robustness of the ANN and the BBN. In this way, we can objectively compare the performance and robustness of the ANN and the BBN based CAD schemes developed in this experiment. A description of the approach, along with the preliminary experimental results derived from three independent databases involving total of 1,557 images, is presented here.

II. Materials and Methods

2.1. Three independent databases

Three independently acquired image databases were used in this study. The first one was used for training the networks. The second was used to evaluate a fitness function in the GA, and the third was used to assess the performance and robustness of the optimized networks. Three databases, which were acquired from three different medical centers, included 545, 579, and 433 images. All of these 1,557 images were digitized in our laboratory using the same laser film digitizer (Lumisys 150) with a pixel size of 100 μm × 100 μm and 12-bit gray-level resolution. The digitized images were then sub-sampled by a factor of four in both directions to generate new images with sizes of approximately 600 × 450 pixels. In each database there is a mixture of images with and without mass regions. All masses were pathology verified. The locations of these verified mass regions in the original film mammograms were identified by expert radiologists in the different medical centers where the mammograms were acquired. Most of the "negative" images in these databases were considered to be difficult controls because they had dense breast parenchyma with highly fluctuated image features. It should be noted that because in this study we were only interested in mass detection, an image without a positive mass was considered as a "negative" image even though it might contain other abnormalities (i.e., microcalcification clusters).

The 1,557 images were individually processed by our multi-layer topographic based CAD scheme which has been described elsewhere [3]. In brief, this scheme has three distinct stages for the identification of suspicious regions. The first stage of dual kernel filtering, subtraction, thresholding, and labeling resulted in the selection of a large number of suspicious regions (approximately 18 regions per image when applied to these image databases). Based on local contrast measurements, the second stage used an adaptive region growth algorithm to define three topographic layers for each suspicious region. In each growth layer, a set of simple intra-layer boundary conditions on the growth ratio and change of shape factor of the region was applied to eliminate a large number of initial suspicious regions (> 80%), which may included both positive and negative regions. Only the regions that successfully pass through three topographic growths were retained as suspicious regions for further classification. After the second stage, the number of suspicious regions (including both positive and negative regions) decreased to 5,560 (approximately 3.6 per image) in these 1,557 images. For each of these remaining regions, a set of image features was automatically computed by our CAD scheme. Using these features, in the third stage of the scheme, different classification tools based on nonlinear multi-layer feature analysis were incorporated to identify positive and negative mass regions. The classification tools that have been tested in our previous studies include a rule-based expert classifier [3], set enumeration trees [8], an ANN [19], and a BBN [20]. In this study, both an ANN and a BBN were used as classification tools. All suspicious regions identified by the second stage of the CAD scheme were included in the experimental databases.

772 of these 5,560 regions depicted verified masses, while 4,788 suspicious regions were actually negative. With the exception of the suspicious regions that matched the verified masses, all other regions that had been identified by our CAD scheme in the second stage as suspicious were determined to be negative. No pathologic verification was available for the negative regions. In summary, there are 288, 172, 312 positive mass regions in these three image databases, respectively. The CAD scheme detected 1,651, 1,876, 1,261 negative (false-positive) regions in these three image databases. A feature vector extracted by the CAD scheme was used to represent each suspicious mass region. This feature vector contains 38 features. Within these features, 32 were computed from the interior of the region and its surrounding background, which were considered to be local features, while the other six were global features

that represent the global tissue patterns of the breast. Definitions of these features and related computational methods using our CAD scheme have been described elsewhere [8,19,20].

2.2. Topology of networks and GA initialization

In these experiments, all 38 features were used to train and test the classifiers. The topologies of the ANN and the BBN are different. The ANN has 38 input neurons, 16 hidden neurons, and one output neuron. The topology of the BBN was similar to the BBN that we have developed and tested for mass detection in our previous studies [20]. Basically, the six global features are located in the top layer of the network and comprise the "parent" nodes to detection node (output for the mass identification). The 32 local features are "child nodes" located in a layer below the detection node. Unlike the ANN where the input features are continuous data (e.g., from 0 to 1), in a BBN, each node must be quantified to a fixed number of exclusive states. The continuous data for these features must be converted to discrete data. In this study, each feature was divided into five discrete states. The methods to convert these features into discrete states have been reported elsewhere [20]. Although each feature vector contains 38 features, many of them might be redundant. The redundant features used in the input nodes of an ANN or a BBN make very little contribution to information but add a lot of noise, which result in poor generalization for the networks. Evan though the topology of a BBN can be interpreted, manual selection of independent and effective features is a difficult task. To find a small number of independent features and eliminate the redundant ones in this feature vector, a genetic algorithm (GA) was used to optimize feature set and topologies of the ANN and the BBN. The GA software, *GENESIS*, was acquired from Prime Time Freeware for AI [21] and used in this study.

A GA solves a complex optimization problem by emulating evolutionary concepts that *only the strongest survives*. A population of possible chromosomes is created, evaluated, recombined, and mutated to generate more and different chromosomes. The best are kept as a basis for evolving better chromosomes. In general, a GA involves the steps of initialization, evaluation, selection, search, and termination [22]. Although the software, *GENESIS*, provides a basic program structure for optimization, users need to determine many detailed parameters and functions, such as encoding, fitness function, and evaluation criterion.

Based on their distributions in our databases, each of the 38 features was normalized to a range between 0 and 1. In the GA, a binary coded chromosome was used. Each feature corresponded to a gene in a chromosome (or a bit of the structure defined in *GENESIS*). In this binary coded chromosome, *1* indicates the presence of a gene (the feature is used as an input node) and *0* indicates its absence (the feature is rejected). In our experiments, all chromosomes have fixed length of 38 (including six global features and 32 local features). The initial population size of the chromosomes was set as 50. In order to incorporate our experience in the feature selection and also to achieve a diverse initial population, about one third of the initial chromosomes were manually selected with a small number of bits of *1* (≤ 10), while the rest of initial population was randomly assigned by GA software. Meanwhile, the crossover rate, the mutation rate, and the generation gap used in the GA were set at 0.6, 0.001, and 1.0, respectively.

2.3. Optimization of feature selection

From initial 38 features, we used GA to select sub-sets of features, $x_i, i = 1,2,\ldots,n$, where $n \leq 38$. The selected features were then connected to the input neurons in the ANN and the probability nodes in the BBN. The number of hidden neurons $(h_j, j = 1,2,\ldots,m)$ in the ANN was determined as half of the input neurons, or $m = (\text{int})(0.5 + n/2)$. There is one neuron in the output layer to represent the result of mass detection. The detailed description of the ANN structure used in our CAD scheme was previously reported [19]. In this study, the number of training iterations was fixed at 1,000. The momentum and learning rate were set as 0.8 and 0.01, respectively. In the BBN, the features selected by the GA were located in the different nodes. If the selected feature was a global feature, it was placed in the top layer, otherwise the feature was connected to one of the "child" nodes in the BBN.

Because a GA is a task independent optimizer, users must provide or define a fitness function and an evaluation criterion, so that the GA has an optimization goal. In the experiments, the fitness function was the receiver operating characteristic (ROC) curve, and the evaluation criterion was the maximum area under the ROC curve (A_z value). ROC curves were generated by the ROCFIT program [23], based on output data from the networks. Once GA selected a chromosome, a set of features was also extracted to be used in the networks. The training database was used to train the ANN by setting its weights or to train the BBN by computing the conditional probability table. After training the networks, the second database (or evaluation database) was used to examine the performance of the networks. The output values of evaluation were directly used as input data to compute a ROC curve. The ROCFIT program produced an A_z value for each selected chromosome. Chromosomes with higher

A_z values have higher priority to be selected to generate new chromosomes in next generation by the GA using the techniques of crossover and mutation. The GA was terminated when it had converged to a maximum A_z value or the searching generation has reached to 50. Two chromosomes that produced highest A_z values for the ANN and the BNN were selected to build an optimal ANN and an optimal BBN, respectively. The performance and robustness of these networks were compared using another independent testing database containing 312 true-positive mass regions and 1,261 false-positive regions. This test database was never involved in any of the optimization processes. Finally, we tested a hybrid classifier, in which feature vectors passed through two networks separately, and the ultimate output was the average score from the outputs of the two networks.

III. Results

Using the same training database and all 38 features to train and test the ANN and BBN, we achieved A_z values of 0.791 ± 0.012 and 0.783 ± 0.011 on the evaluation database involving 172 positive mass regions and 1,876 negative regions for the ANN and BBN, respectively.

Table 1 demonstrates that after GA optimization, the number of features selected in both ANN and the BBN have been significantly reduced. Less than half of the original 38 features were retained for these two networks. Although the features selected in two networks were not exactly the same, the performance levels of the two networks converged to the same level.

Table 1: Optimization results for the ANN and the BBN.

Network	Number of local features	Number of global features	A_z
ANN	12	2	0.866
BBN	14	3	0.868

In an independent test of these optimal networks, the results for the ANN and the BBN remained at the same level. For the ANN, A_z value was 0.847 ± 0.014, and for the BBN, A_z value was 0.845 ± 0.011. Finally, using a hybrid classifier containing both the ANN and the BBN, the A_z value on the test database was increased to 0.859 ± 0.01.

IV. Discussion

Objectively evaluating CAD performance and robustness is a very complicated and difficult task [18]. The performance of a scheme depends on many factors, such as case difficulty in the training and testing databases [24], the size of training database [19], validation methods [25], and the ground truth for the comparison [26]. Basically, the CAD schemes that developed at different institutions and optimized using different databases are not comparable [24]. In this study, we used the same database and the same optimization protocol to train and test two machine learning classifiers, an ANN and a BBN. Hence, the performance of these two networks can be objectively compared. When applied to a new independent database, the performance deterioration of a CAD scheme may be caused by two factors. The first is bias in the training samples due to the limited size of the training set compared to the diverse feature distributions in real clinical testing populations. The second is data over-fitting in certain learning algorithms, which makes classifiers much more sensitive to the noise patterns in the testing samples. The ANN and BBN represent two very typical and popular classifiers used in CAD. The ANN uses a black box "hill-climbing" method to search for the relationship between training samples and classification results. Both the sample bias and the data over-fitting have impact in the robustness of the ANN. The BBN uses Bayesian probability theory to find the optimal relationship between the input features and output results. Although there is no data over-fitting problem in the BBN, bias in the learning samples generates incorrect probability tables that reduce the robustness of the BBN. As a result, the ANN usually outperforms the BBN in training results, but falls behind in testing new cases. In our experiments, several methods have been used to minimize over-fitting during ANN training, which include limiting the number of training iterations and maintaining a large ratio between momentum and learning rate [11]. The training iteration number was limited to 1,000 and the ratio between momentum and learning rate was set at 80.

Although data over-fitting is a potential danger to testing an ANN, by using a GA and setting an appropriate fitness criterion, the impact of over-fitting can be significantly reduced, as shown in this study. In both optimization and independent testing, the ANN and the BBN achieved the same performance level (A_z value), which clearly indicates that, in this experiment, the performance "deterioration" in independent testing is mainly caused by the bias in the training database. There is no significant difference between using an ANN and a BBN in our CAD scheme for mass detection, as long as each network has been properly optimized or trained. This study demonstrated that improving performance and robustness of CAD schemes might be more dependent on feature selection and database diversity than on any particular machine learning or classification algorithm.

V. Acknowledgements

This work is supported in part by the National Cancer Institute under the grants of CA77850 and CA79587, and US Army under grant DAMD17-98-1-8018.

VI. References

1. Vyborny CJ, Giger ML, Computer vision and artificial intelligence in mammography, *Am J. Roentgen* **1994;** 162:699-708.
2. Kegelmeyer WP, Pruneda JM, Bourland PD, Hills A, Riggs MW, Nipper ML, computer-aided mammographic screening for speculated lesions, *Radiology* **1994;** 191:331-337.
3. Zheng B, Chang YH, Gur D, Computerized detection of masses from digitized mammograms using a single image segmentation and a multi-layer topographic feature analysis, *Acad Radiol* **1995;** 2:959-966.
4. Zhang W, Doi K, Giger ML, Nishikawa RM, Schmidt RA, An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms, *Med Phys* **1996;** 23:595-601.
5. Sahiner B, Chan HP, Wei D, Petrick N, Helvie MA, Adler DD, Goodsitt MM, Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue, *Med Phys* **1996;** 23:1671-1684.
6. Li L, Qian W, Clarke LP, Computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms, *Acad Radiol* **1997;** 4:724-731.
7. Polakowski WE, Cournoyer DA, Rogers SK, Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency, *IEEE Trans Med Imaging* **1997;** 16:811-819.
8. Rymon R, Zheng B, Chang YH, Gur D, Incorporation of a set enumeration tree-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection, *Acad Radiol* **1998;** 5:181-187.
9. Cheng HD, Lui YM, Freimanis, A novel approach to microcalcification detection using fuzzy logic technique, *IEEE Trans Med Imaging* **1998;** 17:442-450.
10. Yu S, Guan L, Brown S, Automated detection of clustered microcalcifications in digitized mammogram films, *J. Electronic Imaging* **1999;** 8:76-82.
11. Schalkoff R, Pattern recognition: statistical, structural and neural approaches, John Wiley & Sons, Inc. New York, NY, **1992.**
12. Diederich J, Explanation and artificial neural networks, *Int. J. Man-Machine Stud.* **1992;** 37:335-341.
13. Teach RL, Shortliffe EH, An analysis of physician attitudes regarding computer-based clinical consultation systems, *Comput. Biomed. Res.* **1981;** 14:542-548.
14. Chan HP, Sahiner B, Wagner RF, Petrick N, Effects of sample size on classifier design for computer-aided diagnosis, *Proc SPIE* **1998;** 3338:845-858.
15. Mitchell TM, *Machine learning,* WCB McGraw-Hill, Boston, MA, **1997.**
16. Michie D., Spiegelhalter DJ, Taylor CC, *Machine learning, neural and statistical classification,* Ellis Horwood, New York, NY, **1994.**
17. Kahn CE, Roberts LM, Shaffer KA, Haddawy P, Construction of a Bayesian network for mammographic diagnosis of breast cancer, *Comput. Biol. Med.* **1997;** 27:19-29.
18. Nishikawa RM, Variations in measured performance of CAD schemes due to database composition and scoring protocol, *Proc SPIE on Med Imaging* **1998;** 3338:838-844.
19. Zheng B, Chang YH, Good WF, Gur D, Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme, *Acad Radiol* **1997;** 4:497-502.
20. Zheng B, Chang YH, Wang XH, Good WF, Gur D, Application of a Bayesian belief network in a computer-assisted diagnosis scheme for mass detection, *Proc SPIE on Med Imaging* **1999;** 3661-167.
21. Kantrowitz M, (editor), Prime time freeware for AI, Issue 1-1, Selected materials from the Carnegie Mellon University Artificial Intelligence Repository, **1994.**
22. Hluck G, Genetic algorithms, in chapter 12 of *The handbook of applied expert systems,* edited by Liebowitz J, CRC Press, Boca Raton, FL, **1997.**
23. Metz CE, Kronman HB, Wang PL, Shen JH, ROCFIT: a modified maximum likelihood algorithm for estimating a binormal ROC curve from confidence-rating data, University of Chicago, Chicago, IL, **1985.**
24. Nishikawa, RM, Giger ML, Doi K, Effect of case selection on the performance of computer-aided detection schemes, *Med. Phys* **1994;** 21:265-269.
25. Tourassi GD, Floyd CE, The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis, *Med. Decis. Making* **1997;** 17:186-197.
26. Kallergi M, Carney GM, Gaviria J, Evaluating the performance of detection algorithms in digital mammography, *Med. Phys.* **1999;** 26:267-275.

Proceedings of the IASTED International Conference
On Artificial Intelligence and Soft Computing
August 9-12, 1999 in Honolulu, Hawaii, USA

Applying A Genetic Algorithm for the Improvement of Decision Making in Medical Imaging Diagnosis [1]

Bin Zheng, Walter F. Good, Yuan-Hsiang Chang, Xiao-Hui Wang
A437 Scaife Hall, University of Pittsburgh, PA 15261, USA
bzheng@radserv.arad.upmc.edu

ABSTRACT

This study is to investigate the effectiveness of using genetic algorithm (GA) to optimize the feature selection and improve the performance of decision making in computer-assisted diagnosis (CAD) schemes of mammography. Three databases involving 1,557 images were used. A CAD scheme was applied to search for initial suspicious mass regions in images and extracted 32 features from each region. Two different classifiers, an artificial neural network (ANN) and a Bayesian belief network (BBN), were then used to identify positive and negative mass regions. Using GA optimization, about half of 32 features was eliminated from both the ANN and the BBN to obtain optimal performance. Although GA selected different features for the ANN and BBN, using ROC analysis, two networks yielded similar performance. Compared to the networks using 32 features, the optimal ANN and BBN obtained better performance. The area under ROC (A_z value) was increased from 0.81 to 0.88 and 0.79 to 0.88 in the ANN and BBN, respectively. The robustness of both networks was tested using an independent database which produced A_z values higher than 0.86. This study demonstrated that (1) the GA could provide an effective approach to optimizing CAD schemes, and (2) the performance of the CAD schemes depended more on feature selection and database diversity than on the particular classification method.

Key Words: Genetic algorithm, Computer-assisted diagnosis, Artificial neural network, Bayesian belief network, Medical decision making.

1. INTRODUCTION

There is a rapidly growing interest in developing computer-assisted diagnosis (CAD) schemes to provide assistance with decision making in medical image diagnosis. For example, in mammography, after very intensive investigation by a large number of research groups around the world for more than a decade, many CAD schemes have been developed [1-10]. For mass detection, CAD schemes usually follow three steps in identifying positive and negative regions. In the first step, the CAD schemes use different image segmentation methods to filter and select initial suspicious regions. In the second step, CAD schemes compute and extract a set of features from each suspicious region. Then, in the third step, a classifier (i.e., an artificial neural network (ANN)) is used to identify positive and negative regions based on a subset of the extracted features. Due to the complexity of imaged breast tissue patterns, visually defining effective features to identify positive and negative mass regions is extremely difficult. Thus, a large number of image features are initially extracted from each suspicious mass region by many CAD schemes. For example, the number of initially selected features can be as much as 587 [7]. Actually, in a small suspicious region, only small number of features has independent values, while others are redundant. Because of the noise in the database, increasing the number of redundant features can significantly deteriorate robustness of a classifier [11]. As a result, optimally selecting features is very important for achieving good performance and robustness of CAD schemes.

Many feature selection methods have been tested in CAD schemes, such as empirical histogram and correlation analysis [4,11], stepwise feature analysis [5], and genetic algorithms (GA) approach [7]. Because most of the features extracted by CAD schemes are not visible or understandable to human observers, machine learning methods may be more effective and efficient for feature selection for CAD schemes. Among these methods, GA techniques have attracted much attention. Using a large number of random seeds spread over the feature space, a GA has capability of parallel processing and searching for a subset of features that can generate an optimal classification result based on its evaluation criterion. Although GAs have been employed for feature selection for CAD schemes by other researchers [7,12], it is not clear whether the GA selected the optimal feature set, because GA is susceptible to problems of the "hill-climbing" process, and it does not guarantee finding global maxima. In this study, we investigated the effectiveness of GA in feature selection by comparing the results with those obtained using a progressive round-off searching method [13]. Although it is not as efficient, the

progressive round-off method can guarantee finding "close-to-best" performance.

Specifically, in this study, GA was used to optimize the feature set and topology of an ANN and a BBN (Bayesian belief network) incorporated in our CAD scheme for mass detection. A description of the approach, along with the experimental results with three independent databases involving total of 1,557 digitized mammograms, is presented here.

2. MATERIALS AND METHODS

2.1. Databases:

Three independently acquired image databases from three different medical centers were used in this study. The first database was used to train the weights in the ANN or compute the conditional probabilities in the BBN. The second database was used to evaluate the performance of two networks. The third database served as an independent testing database to examine robustness of the CAD scheme incorporating the ANN or the BBN after GA optimization. There are 545, 433, and 579 film mammograms in these three databases, respectively. All 1,557 mammograms were digitized in our laboratory using the same laser film digitizer with a pixel size of 100 μm × 100 μm and 12-bit gray-level resolution. The digitized images were then sub-sampled by a factor of four in both directions to generate new images of approximately 600 × 450 pixels. In each database there is a mixture of images with and without positive mass regions. All positive masses were pathology verified. The locations of the visible mass regions in the original film mammograms were marked by expert radiologists in the three medical centers where these mammograms were acquired. Most of the "negative" images in the databases were considered to be difficult controls, because they had dense breast parenchyma with highly fluctuated image features.

Each image was then processed by a multi-layer topographic based CAD scheme developed in our laboratory [3]. In brief, this scheme has three distinct stages for the identification of suspicious masses. The first stage of dual kernel filtering, subtraction, and labeling resulted in the selection of a large number of suspicious regions (approximately 20 regions per image when applied to these image databases). Based on local contrast measurements, the second stage used an adaptive region growth algorithm to define three topographic layers for each suspicious region. In each growth layer, a set of simple intra-layer boundary conditions on region growth ratio and shape factor of the region was applied to eliminate many initial suspicious regions (> 80%), which included both positive and negative regions. Only the regions that successfully passed three topographic region growths were retained for further classification. After the second stage, the number of suspicious regions (including both positive and negative regions) decreased to 6,774 (approximately 4.35 per image) in these 1,557 images. For each of the remaining regions, a set of image features was automatically computed by the scheme. Using these features, the third stage of the CAD scheme used a classification tool based on nonlinear multi-layer feature analysis to identify positive and negative mass regions. The classification tools that have been tested in our CAD studies include a rule-based expert system [3], set enumeration trees [9], an ANN [14], and a BBN [15]. In this study, two classifiers, an ANN and a BBN, were tested and compared. All of the 6,774 suspicious regions identified by the second stage of the CAD scheme were included in the experimental databases.

742 of these 6,774 regions depicted verified masses, while 6,032 suspicious regions were actually negative. With the exception of the suspicious regions that matched the verified masses, all other regions that identified as suspicious by our CAD scheme in the second stage were determined to be negative. In summary, the training database included 288 mass regions and 2,202 actually negative regions. The evaluating database had 304 mass regions and 1,586 negative regions, and the database for testing the robustness of the CAD scheme consisted of 150 mass regions and 2,252 negative regions. The image feature distributions (or conspicuity) of these positive and negative mass regions have been reported elsewhere [16]. For each suspicious mass region, a vector of 32 features was extracted by the CAD scheme. Within these features, 27 were computed from interior of the region and its surrounding background, which were considered as local features, while the remaining 5 were global features that represent the location of the region in the image and overall tissue patterns of the breast. The definitions and methods used to compute these features by our CAD scheme have been described elsewhere [14-16].

2.2. Progressive round-off search:

The performance of the ANN and the BBN is evaluated using ROC (receiver operating characteristic) methodology. The outputs of sets of testing samples from the classifiers are evaluated by a standard ROC analysis program [17]. A larger area under ROC curve (or the higher A_z value) indicates better performance of the classifiers. Before we tested the effectiveness of the GA in optimizing the feature set for our CAD scheme, we defined a "close-to-best" reference that was selected from the initial 32 features. The most reliable approach to finding a best feature set (the global maxima in the feature space) would be a complete permutation search, as we have previously demonstrated [18]. However, due to limitations of computing power, applying a complete permutation search to these 32 features is computationally impractical. An alternative defining a "close-to-best" feature set is to use a progressive round-off search. To search for an optimal set of features from N extracted features, the progressive round-off method starts from m

features ($m<N$) that are previously identified as belonging to the "close-to-best" feature set. From these m features, the approach adds one of the remaining features to the feature set. After finding the highest performance involving $m+1$ features, the scheme fixes these $m+1$ features and searches for an additional feature. The process continues recursively until performance can not be further improved by adding a new feature [13]. Our previous study [18] demonstrated that this progressive round-off method could find a "close-to-best" feature set for a BBN incorporated in a CAD scheme. An exhaustive permutation method was used to search for initial 10 optimal features in this experiment. Then the progressive round-off method was applied to find the remaining features to be included in the "close-to-best" reference.

2.3. Initialization of GA:

The GA software, *GENESIS* [19], was used in this experiment to search for the optimal features for the ANN and the BBN. Each feature corresponded to a gene in a chromosome. Thus, all chromosomes have fixed length of 32 (including 5 global and 27 local features). Binary coding was used to create a chromosome, with *1* indicating the presence of a gene and *0* indicating absence of a gene. Other initial parameters required in the software were selected and adjusted based on the experimental results. The initial population size of chromosomes was set at 50. In order to incorporate our experience in the feature selection and also to achieve a diverse initial population, about one third of initial chromosomes were manually selected with small number of bits of *1* (≤ 7), while the rest of initial population was randomly assigned by GA software. The crossover rate, the mutation rate, and the generation gap were set at 0.6, 0.001, and 1.0, respectively.

2.4. Optimization of feature set using GA:

Two classification networks, ANN and BBN, were used to identify positive masses. The detailed ANN and BBN topologies used in our CAD studies have been reported before [14,15]. Because GA is a task independent optimizer, users must define a fitness function and an evaluation criterion reflecting the optimization goal of the GA. In this experiment, the fitness function was A_z, the area under ROC curve, as determined by the ROC analysis. In each GA search, a set of features represented by a selected chromosome was extracted from the training database. The training samples were used to train the weights connecting the neurons in the ANN or compute the conditional probability table in the BBN. The second database was then used to examine the performance of the networks. The evaluation criterion of the GA was the highest A_z value of this database. The chromosomes with higher A_z values had higher probabilities of being selected in generating new chromosomes for the next generation by GA using the methods of crossover and mutation. The GA was terminated when it converged to the highest A_z value.

Since the effectiveness of the GA depends on the selection of many initial parameters, we first used GA to optimize feature selection in the BBN, so that we can test the effectiveness of GA by comparing the result with that obtained from the progressive round-off approach. In the BBN, five global image features were used as fixed parent nodes of the mass detection node. The child nodes of the mass detection were selected from 27 local image features. The detailed description of the BBN for mass detection has been reported elsewhere [15]. After the best initial parameters were determined in the GA, the GA was applied to search for the optimal feature set for the ANN. The selected features ($n < 32$) by the GA were used as input neurons in the ANN and the number of hidden neurons was set at half the number of input neurons. Weights in the ANN were trained using the training database. In our experiments, several methods have been used to minimize over-fitting during ANN training, which include to limiting the number of training iterations and maintaining a large ratio between the momentum and learning rate. The number of training iterations was fixed at 1,500, while the momentum and learning rate were set as 0.8 and 0.01, respectively. Ratio between momentum and learning rate was 80. The top five chromosomes (feature sets) selected by the GA search for the ANN and the BBN were recorded. They were used to demonstrate the optimal performance achieved by the ANN and the BBN. The robustness of two networks was also examined using an independent testing database, which had not previously been involved in GA optimization process.

3. RESULTS

Using all 32 features as input nodes of the two networks, the A_z values were 0.813 ± 0.015 and 0.794 ± 0.012 from the evaluating database involving 304 positive mass regions and 1,586 negative regions in the ANN and the BBN, respectively. The progressive round-off method selected 15 features that could achieve the highest A_z value of 0.876 ± 0.011 in the BBN. Using the initial parameters for the GA, the highest A_z value obtained in the optimal BBN search was 0.859 ± 0.009 using 17 features. Although the performance was better than using all of the 32 features, GA did not find the best feature set. In the second test, the population size of the GA was increased from 50 to 100. The best chromosome found by GA represented the same "close-to-best" set of features determined in the progressive round-off method.

Then, the GA was used to find an optimal feature set for the ANN. Tables 1 and 2 list the top five optimal feature subsets and their corresponding A_z values for

training, evaluating, and testing databases in the BBN and the ANN.

Table 1: Top five optimal feature subsets in the BBN.

Feature set	# of feature	Training database	Evaluation database	Testing database
1	15	0.906	0.873	0.865
2	16	0.901	0.871	0.863
3	14	0.904	0.873	0.863
4	**15**	**0.905**	**0.876**	**0.868**
5	17	0.905	0.872	0.864

Table 2: Top five optimal feature subsets in the ANN.

Feature set	# of feature	Training database	Evaluation database	Testing database
1	12	0.891	0.872	0.832
2	10	0.914	0.873	0.818
3	**16**	**0.906**	**0.879**	**0.862**
4	13	0.909	0.873	0.841
5	16	0.921	0.873	0.831

The average of top five A_z values for both the ANN and the BBN was increased to the same level (about 0.88). When using the optimized networks to test an independent database, the A_z values in five ANN tests demonstrated larger fluctuation with maximum difference of 0.044. The five A_z values in the BBN tests were very stable (with maximum differences of 0.002). Figure 1 demonstrates two ROC curves representing the best testing performance in the ANN and the BBN.

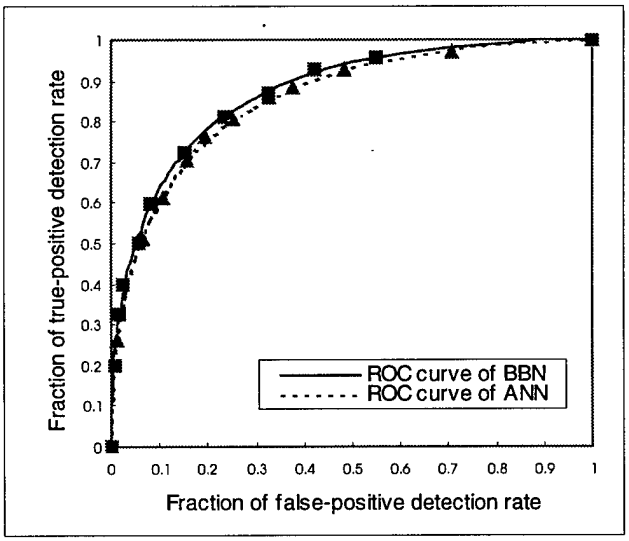

Figure 1: ROC curves comparing the two highest performing BBN and ANN on the testing database. The A_z values are 0.868 ± 0.009 for the BBN and 0.862 ± 0.012 for the ANN.

4. DISCUSSIONS

The performance deterioration in the CAD schemes, when they are applied to new independent databases, may be caused by two factors. The first is a bias in training samples, due to the limited size of training set and the diverse feature distribution in real clinical testing populations. The second is over-fitting in certain training algorithms. Over-fitting makes a machine learning method more sensitive to the noise patterns in the testing samples. The ANN and the BBN represent two popular classifiers, or decision making tools, used in medical image diagnosis. The ANN uses a simple heuristic search to find relationship between the training samples and classification results. Both the sample bias and over-fitting have impact in the robustness of the ANN. The BBN uses Bayesian probability theory to find the optimal relationship between the input features and output results. Although there is no over-fitting problem in the BBN, the bias in the learning samples can generate incorrect probability tables, which reduce the robustness of the BBN. This study demonstrated that optimization of the feature set (or input nodes) used in the ANN and the BBN played an important role to in improving CAD performance and robustness.

Feature extraction can be considered as a form of data compression that removes irrelevant information and preserves relevant information from the raw data. Usually, large number of features is nitially extracted from a suspicious mass region in mammogram. Most of these features are redundant. The incorporation of an optimal set of features in an ANN or a BN presents issues that are similar to the typical signal-to-noise ratio problem. In real application of medical image processing, every feature contains both information (signal) and noise. The redundant features used in the input nodes of the ANN or the BBN make very little contribution to the information but add a lot of noise. An important lesson about generalization (or robustness) of a supervised machine learning classifiers can be learned from statistics; too many free parameters results in over-fitting. A curve fitted with too many parameters follows all small details or noise but is very poor for interpolation and extrapolation [20]. The same is true for the ANN and the BBN. Too many weights (or nodes) in a network give poor generalization. Therefore, optimization of feature sets for medical image diagnosis involving an ANN or a BBN is important. Many methods of feature selection or optimization have been reported in medical image processing. Most of these rely heavily on input or judgement from users' empirical knowledge, which is very difficult. The advantage of using a GA is that users do not need to have much pre-knowledge about the features. A GA has the ability to find a "close-to-best" feature set for use in different machine learning algorithms, if the proper initial parameters are chosen. Compared to the progressive round-off search methods, A GA is much more efficient. This experiment demonstrated

that using appropriate initial parameters and a fitness criterion, the GA could be used as an effective approach for optimizing feature subset and network topology for decision making in the CAD of mammography.

Although over-fitting is a potential danger in training an ANN, this study also found that by using a GA optimization technique with an appropriate fitness criterion, the impact of over-fitting could be significantly reduced in the ANN. After separating the databases used for training or computing the weights in the networks and for optimizing the fitness evaluation in the GA, the performance and robustness of both the ANN and the BBN on a new independent testing database converged to the same level. Therefore, the improving the performance and robustness of CAD schemes may be more dependent on feature selection and database diversity than on any particular machine learning or classification paradigm. Although this experiment was applied to the CAD of mammography, the conclusions obtained could also be applied to other schemes used for computer-assisted diagnosis on other kinds of medical images.

5. ACKNOWLEDGMENTS

This work was supported in part by National Cancer Institute (USA) under grants CA77850, CA82912, and US Army under grant DAMD17-98-1-8018.

6. REFERENCES

1. C.J. Vyborny, and M.L. Giger, Computer vision and artificial intelligence in mammography, *Am J. Roentgen*, 162, 1994; 699-708.
2. W.P. Kegelmeyer, J.M. Pruneda, and M.L. Nipper, Computer-aided mammographic screening for speculated lesions, *Radiology*, 191, 1994, 331-337.
3. B. Zheng, Y.H. Chang, D: Gur, Computerized detection of masses from digitized mammograms using a single image segmentation and a multi-layer topographic feature analysis, *Acad Radiol,* 2, 1995, 959-966.
4. H.D. Li, M. Kallergi, L.P. Clarke, V.K. Jain, and R.A. Clark, Markov random field for tumor detection in digital mammography, *IEEE Trans Med Imaging,* 14, 1995, 565-576.
5. H.P. Chan, D. Wei, M.A. Helvie, B. Sahiner, D.D. Adler, and M.M. Goodsitt, Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture space, *Phys. Med. Biol.*, 40, 1995, 857-876.
6. W. Zhang, K. Doi, M.L. Giger, R.M. Nishikawa, and R.A. Schmidt, An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms, *Med Phys,* 23, 1996, 595-601.
7. B. Sahiner, H.P. Chan, D. Wei, and M.M. Goodsitt, Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue, *Med Phys,* 23, 1996, 1671-1684.
8. W.E. Polakowski, D.A. Cournoyer, S.K. Rogers, Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency, *IEEE Trans Med Imaging,* 16, 1997, 811-819.
9. R. Rymon, B. Zheng, Y.H. Chang, and D. Gur, Incorporation of a set enumeration tree-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection, *Acad Radiol,* 5:181-187, 1998.
10. H.D. Cheng, Y.M. Lui, R.I. Freimanis, A novel approach to microcalcification detection using fuzzy logic technique, *IEEE Trans Med Imaging,* 17, 1998, 442-450.
11. Y. Wu, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, and C.E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology,* 187:81-87, 1993.
12. M. Anastasio, H. Yoshida, R. Nagel, R.M. Nishikawa, and K. Doi, A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms, *Med. Phys.*, 25, 1998, 1613-1620.
13. K. McAloon, and C. Tretkoff, *Optimization and computational logic* (New York; John Wiley and Sons, Inc., 1996).
14. B. Zheng, Y.H. Chang, W.F. Good, and D. Gur, Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme, *Acad Radiol,* 4, 1997, 497-502.
15. B. Zheng, Y.H. Chang, X.H. Wang, W.F. Good, and D. Gur, Application of a Bayesian belief network in a computer-assisted diagnosis scheme for mass detection, *PROC SPIE imaging processing,* 1999, 3661-167.
16. B. Zheng, Y.H. Chang, W.F. Good, and D. Gur, Assessment of mass detection using tissue background information as input to a computer-assisted diagnosis scheme, *PROC SPIE Imaging processing,* 3338, 1998, 1547-1555.
17. C.E. Metz, H.B. Kronman, P.L. Wang, and J.H. Shen, *ROCFIT: A modified maximum likelihood algorithm for estimating a binormal ROC curve from confidence-rating data,* (Chicago; University of Chicago, 1985).
18. B. Zheng, Y.H. Chang, X.H. Wang, W.F. Good, and D. Gur, Feature selection for computerized mass detection in digitized mammograms using a genetic algorithm, *Acad Radiol,* 1999, in press.
19. M. Kantrowitz, *Prime time freeware for AI, Issue 1-1*, (Pittsburgh; Carnegie Mellon University, 1994).
20. J. Hertz, A. Krogh, R.G. Palmer, *Introduction to the theory of neural computation,* (Redwood City, CA; Addison-Wesley Co., 1991).